A Benchmark for the Use of Topic Models for Text Visualization Tasks

Daniel Atzberger*
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

Tim Cech*
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

Willy Scheibel
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

Daniel Limberger
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

Matthias Trapp
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

Jürgen Döllner
Hasso Plattner Institute,
Digital Engineering Faculty,
University of Potsdam, Germany

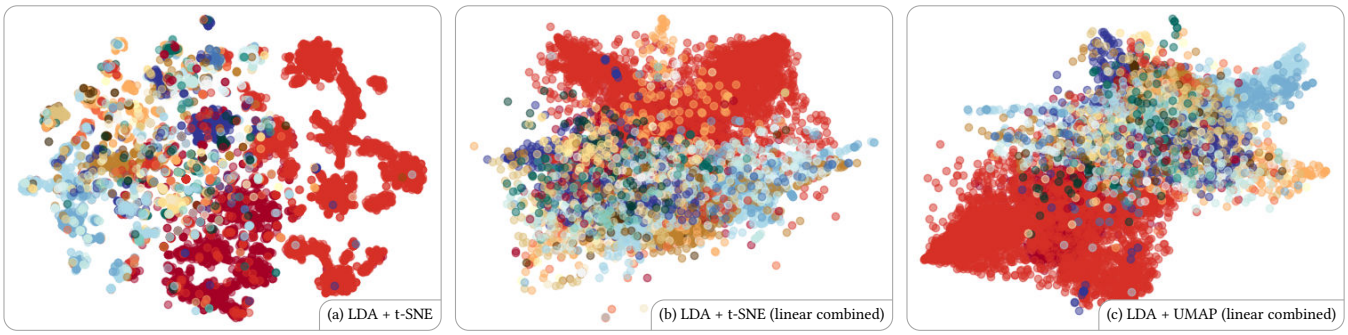

Figure 1: A visual comparison of three layouts for the *Reuters* dataset. Each point presents one document; the color represents the category associated with the document. The first layout (a) results from the application of LDA together with t-SNE; the second one (b) from LDA and t-SNE applied on the topics; the third one (c) from applying LDA together with UMAP on the extracted topics.

ABSTRACT

Based on the assumption that semantic relatedness between documents is reflected in the distribution of the vocabulary, topic models are a widely used class of techniques for text analysis tasks. The application of topic models results in concepts, the so-called topics, and a high-dimensional description of the documents. For visualization tasks, they can be projected onto a lower-dimensional space using dimensionality reduction techniques. Though the quality of the resulting point layout mainly depends on the chosen topic model and dimensionality reduction technique, it is unclear which particular combinations are suitable for displaying the semantic relatedness between the documents. In this work, we propose a benchmark comprising various datasets, layout algorithms and their hyperparameters, and quality metrics for conducting an empirical study.

*Both authors contributed equally to this work

VINCI'22, August 16–18, 2022, Chur, Switzerland

ACM ISBN 978-1-4503-9806-0/22/08.
https://doi.org/10.1145/3554944.3554961

CCS CONCEPTS

• **Human-centered computing** → **Visual Analytics**.

KEYWORDS

Text Visualization, Topic Modeling, Dimensionality Reduction Techniques

ACM Reference Format:
Daniel Atzberger, Tim Cech, Willy Scheibel, Daniel Limberger, Matthias Trapp, and Jürgen Döllner. 2022. A Benchmark for the Use of Topic Models for Text Visualization Tasks. In *15th International Symposium on Visual Information Communication and Interaction (VINCI'22), August 16–18, 2022, Chur, Switzerland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3554944.3554961

1 INTRODUCTION

One of the most fundamental questions related to the analysis of text collections is the clustering of documents according to their semantics. In addition to pure text clustering techniques, various visualization approaches have been developed to derive these semantic distributions within a given corpus from a spatial representation by exploiting human perceptual capabilities. Most such visualizations are based on a two-dimensional or three-dimensional point layout, where a point represents each document. The pairwise Euclidean distance between two points reflects the semantic similarity between the documents. This point layout can be used as

a scatter plot that can further be enriched by mapping data of the documents, e.g., their word count or their group label in the case of labeled documents, onto visual variables of the geometric objects. For example, a contour plot can be derived from a two-dimensional scatter plot, thus displaying the word count [11], or 3D glyphs can be placed on a reference plane, thus offering a large number of mappings [1, 2]. Here and in the following, when speaking about a (point) layout, we always refer to the resulting scatter plot of a layout algorithm, which is a combination of a topic model and a subsequent dimensionality reduction technique. In the particular case of text documents, we refer to a semantic layout (algorithm) for displaying semantic relatedness between the documents.

All approaches assume that the semantic similarity of documents can be inferred from the distribution of words across documents. In this context, the description of the semantic structure of a set of documents, the so-called corpus, is usually based on the so-called Bag-of-Words (BOW) assumption, i.e., one neglects the ordering of the words and stores only the frequencies within a document. The resulting description of the corpus is called the vector space model, i.e., the corpus is stored in a document-term-matrix, whose entries store the frequency of a word in the respective document, and each document is represented by one row in the matrix. A lower-dimensional representation of the corpus – a point layout – is then derived using a dimensionality reduction technique combined with a previously trained topic model. However, the quality of the generated point layout depends on the combination of topic models and dimensionality reduction techniques and the chosen hyperparameters. Figure 1 shows the results of three dimensionality reduction techniques applied with the topic model approach LDA to the same dataset, which differ strongly from each other. Although various combinations of topic models and dimensionality reduction techniques have been proposed, it is unclear which layout algorithm is best suited for visualizing semantic relatedness.

Inspired by previous works on quantitative assessments of dimensionality reduction techniques for visualization tasks [7], we propose a benchmark $\mathcal{B} = (\mathcal{D}, \mathcal{L}, \mathcal{Q})$ to investigate the effectiveness of topic models for visualizing the semantic relatedness between documents. Our benchmark consists of three parts: A set of text corpora $\mathcal{D}$ whose elements are collections of documents, the set $\mathcal{L}$ of layout algorithms that map each document within a corpus onto a two-dimensional plane, and the set $\mathcal{Q}$ of quality metrics that measure how well the point layout algorithm preserves local and global structures.

2 RELATED WORK

Our presentation of the related work comprises two parts, (1) approaches for visualizing text corpora utilizing topic models and dimensionality reduction techniques, and (2) evaluations of the use of dimensionality reduction techniques.

Wise et al. were one of the first who proposed an algorithm for visualizing text corpora using a spatial representation [20]. Based on a high-dimensional description derived from statistical properties of the documents, a lower-dimensional point layout is derived by applying Principal Component Analysis (PCA) or Multidimensional Scaling (MDS) as a subsequent dimensionality reduction technique.

By applying Self-organizing Maps (SOM) on the document-term-matrix, Skupin presented a layout algorithm for generating a 2D visualization for publication abstracts [17]. Their final visualization uses cartographic metaphors to generate a map-like information landscape.

Gansner et al. applied LDA to a text corpus to derive a graph structure that facilitates the similarity between documents [8]. Their visualization approach can handle streamed data using a dynamic graph layout algorithm.

In addition, Le and Lauw developed a generative model, similar to LDA for text corpora [12]. However, in contrast to the generative model underlying LDA, their model infers coordinates for the documents directly from the term-document-matrix rather than deriving it from a subsequent dimensionality reduction.

Kucher et al. visualized various statistics of proceedings of VINCI for a total of nine years [10]. In doing so, the authors applied LDA together with t-SNE to create a two-dimensional layout for content analysis.

Caillou et al. presented *Cartolabe*, a web-based system for visualizing text corpora [4]. First, the documents are placed on a two-dimensional plane using LDA or Latent Semantic Indexing (LSI) together with UMAP. Subsequently, a meaningful label is extracted for each region.

Atzberger et al. applied LDA to capture the semantic structure of software projects [1]. In their layout algorithm, the topic-word distributions are first projected using MDS applied on the dissimilarity matrix capturing the pairwise Jensen-Shannon distances, as presented with the LDAvis approach [16]. The positions of the documents are then aggregated as convex linear combinations according to their document-topic distributions. In a later work, the authors created a *KnowhowMap*, capturing the similarity between developers in a 2.5D visualization, based on the proposed layout algorithm [2].

The different dimensionality reduction techniques differ in a number of characteristics, including preservation of local or global structure in the high-dimensional data. To capture their characteristics and applicability, the techniques were measured and evaluated in prior work. As such, Espadoto et al. presented a benchmark for studying their for visualization tasks [7]. The authors analyzed 18 datasets categorized into tabular, image, and text datasets, and a total of 44 projection techniques. By evaluating the results using seven quality metrics, the authors draw several conclusions on the practical use of dimensionality reduction techniques, e.g., what projection techniques are suitable for which data type, the sensitivity of the results on the parameters, or questions related to the performance. The benchmark of Espadoto et al. exceeds previous studies in their size and focuses more on practical aspects than theoretical ones.

3 BENCHMARK

The goal of a semantic layout algorithm is to place documents that share large parts of their semantic structure nearby. For comparing semantic layout algorithms for text corpora, we present a benchmark $\mathcal{B} = (\mathcal{D}, \mathcal{L}, \mathcal{Q})$ consisting of (1) a set of datasets $\mathcal{D}$, (2) a set of layout algorithms $\mathcal{L}$, and (3) a set of quality measures $\mathcal{Q}$ for quantifying the results. Our work is mainly inspired by the work

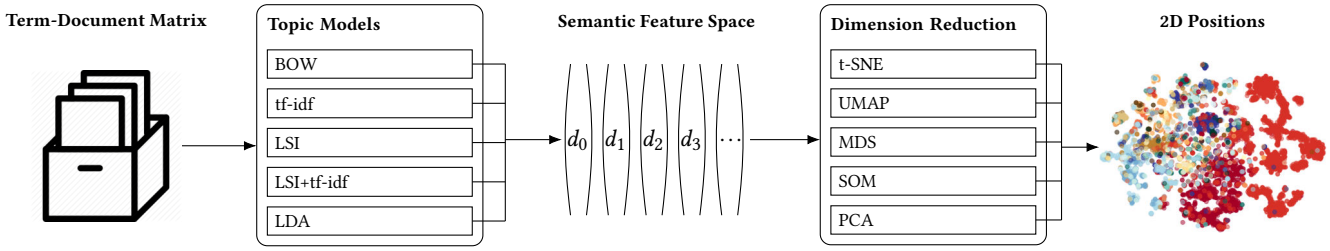

Figure 2: Pipeline for generating a semantic point layout for a text corpus, where the dissimilarity between the topics is not taken into account.

of Espadoto et al., who proposed and evaluated a benchmark for evaluating the effectiveness of various dimensionality reduction techniques for visualization tasks [7]. Even though the authors consider text data as a main data category, they do not take topic models as a modelling step into account.

3.1 Datasets $\mathcal{D}$

Our set of datsets $\mathcal{D}$ consists of text corpora, i.e., each element $D \in \mathcal{D}$ is itself a set of documents $D = \{d_1, \ldots, d_m\}$. For each corpus, we undertake its underlying vocabulary several preprocessing steps to reduce the size of the vocabulary and to remove words that carry no meaning. Examples of such preprocessing steps include the removal of stop words, the lemmatization of the words, and corpus-specific steps, e.g., the removal of email headers. After preprocessing a corpus, D can be viewed as a matrix of size $m \times N$, where N denotes the size of the vocabulary, i.e., the number of words within the corpus. The cell at position (i, j) denotes the number of occurrences of the j-th word in the i-th document. After preprocessing, $\mathcal{D}$ comprises the following elements:

(1) **20 Newsgroup**: A set of 18827 documents categorized into 20 classes with a total vocabulary size of 23959.
(2) **Reuters**: Our chosen subset of the Reuters dataset contains 11367 documents, as not all documents in the original dataset come with a class label. Our subset comes with 69 class labels and a vocabulary size of 10391.
(3) **agnews**: A corpus of 127599 documents divided into four groups and a vocabulary size of 20860.
(4) **blogs:** A corpus of 99999 documents from 40 categories with a total vocabulary size of 35699.
(5) **smsspam**: A set of 5574 messages with an assigned binary variable. The total size of the vocabulary is 3599.
(6) **hatespeech**: A corpus of 24783 documents from three categories and a total vocabulary size of 8176.
(7) **imdb**: A set of 49999 documents from two categories and a vocabulary of size 30354.

3.2 Layout Algorithms $\mathcal{L}$

Various semantic layout algorithms for corpora arising natural language have been proposed, specifying concrete processing techniques of the term-document-matrix and a subsequent dimensionality reduction technique. However, all proposed layout algorithms can be deduced from a general model by specifying (1) the topic model for capturing the semantic structure of D, (2) whether the topic-term distribution is taken into account or not, (3) the applied dimensionality reduction technique to generate a 2D or 3D result. An overview of the pipeline for generating a two-dimensional layout for displaying semantic relatedness between documents is shown in figure 2.

A dimensionality reduction technique either requires a set of data points or a square matrix capturing the pairwise dissimilarities between the data points. In both cases, a description of the documents as feature vectors is required. Various learning algorithms have been proposed for comparing documents on a semantic level based on the distribution of the words among the documents. In our considerations, we focus on the following techniques:

(1) **Pure Bag-of-Words (BOW):** In the pure BOW model, the corpus D is undertaken no further processing, i.e., each document is described by its corresponding row in the document-term-matrix. The documents are then compared using the cosine similarity.
(2) **BOW with tf-idf weighting:** In the pure BOW model, more frequently occurring words have a higher impact on the comparison than less frequent words. However, this assumption often is not valid, e.g., words that occur very often among the entire corpus. The tf-idf weighting scheme addresses this issue by replacing the frequency of the terms with the product of the term frequency and inverse document frequency.
(3) **LSI** is a topic model based on Singular Value Decomposition [6]. The algorithm extracts so-called topics, which are vectors of length given by the vocabulary size. Each topic vector entry describes a word's impact on the topic. The documents are modeled as vectors whose length corresponds to the number of topics. Before applying LSI, the term-document-matrix is sometimes undertaken the tf-idf weighting scheme.
(4) **LDA** is a probabilistic topic model, which assumes a generative process underlying a corpus [3]. By examining patterns of co-occurring words, it extracts topics as multinomial distributions over the vocabulary, where the number of the topics is a hyperparameter of the model. The documents are described as distributions over the extracted topics and are usually compared using the Jensen-Shannon distance.

Besides a document-topic distribution, LDA results in a topic-term distribution that captures latent concepts within a corpus. In most works, the fact that two topics can have a varying degree of similarity is neglected. Therefore the document-topic distribution is reduced to generate a low-dimensional plot. However, an alternative method, taking the varying degree of dissimilarity of topics into

account, was presented by Atzberger et al. [1]. In this variant, the topic-term distributions are reduced in their dimension, and the position of documents is then computed as a convex linear sum according to their document-topic distribution.

Numerous dimensionality reduction techniques have been developed to capture local and global structures of high-dimensional datasets. In our considerations, we select a subset among all those techniques that have either shown promising results in the study [7] for visualizing text data or have been applied in existing semantic layout approaches. In concrete, we examine the following dimensionality reduction techniques: t-SNE [19], UMAP [13], Projection by Clustering (PBC) [15], MDS [5], SOM [9], Interactive Document Maps (IDMAP) [14], Local Linear Coordination (LLC) [18].

3.3 Quality Metrics Q

The effectiveness of a layout algorithm to capture the semantic relatedness between the documents can be evaluated using different quality metrics [7]. For our experiments, we choose the following quality metrics:

(1) **Trustworthiness** and **Continuity** for measuring the quality of local structures in the lower-dimensional space.
(2) **Normalized stress**, which measures how well distances in the semantic feature space are preserved in its lower-dimensional representation.
(3) **3-Neighborhood hit**, which measures how well labels are respected in the lower-dimensional space.

4 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a benchmark that can be used to evaluate the effect of topic models and dimensionality reduction techniques for creating 2D or 3D spatial representations of text corpora. Our first experiments indicate that the choice of layout algorithm can significantly impact the resulting visualization, thus motivating a more detailed study of the problem. Based on the current results, we suspect that the best results can be achieved using LDA as the topic model and applying t-SNE on the resulting topic-word distributions. The position of the documents is then computed as a linear combination according to their document-topic distribution.

Our work is still at an early stage. By statistical investigation of the results, we want to answer whether topic models are beneficial for visualizing text corpora or not and which combinations of topic models and dimensionality reduction techniques result in high-quality semantic layouts. Besides quantifying the effectiveness of preserving local and global structures, it would also be beneficial to evaluate the resulting layout concerning its capabilities to support user perception. Furthermore, we want to evaluate to what extent an optimized layout improves the various visualization techniques for different user tasks.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers for their valuable feedback to improve this article. This work is part of the "Software-DNA" project, which is funded by the European Regional Development Fund (ERDF or EFRE in German) and the State of Brandenburg (ILB). This work is part of the KMU project "KnowhowAnalyzer" (Förderkennzeichen 01IS20088B), which is funded by the German Ministry for Education and Research (Bundesministerium für Bildung und Forschung).

REFERENCES

[1] Daniel Atzberger, Tim Cech, Merlin de la Haye, Maximilian Söchting, Willy Scheibel, Daniel Limberger, and Jürgen Döllner. 2021. Software Forest: A Visualization of Semantic Similarities in Source Code using a Tree Metaphor. In *Proc. 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 3 (IVAPP '21)*. INSTICC, SciTePress, 112–122. https://doi.org/10.5220/0010267601120122

[2] Daniel Atzberger, Tim Cech, Adrian Jobst, Willy Scheibel, Daniel Limberger, Matthias Trapp, and Jürgen Döllner. 2022. Visualization of Knowledge Distribution across Development Teams using 2.5D Semantic Software Maps. In *Proc. 13th International Conference on Information Visualization Theory and Applications (IVAPP '22)*. INSTICC, SciTePress, 210–217. https://doi.org/10.5220/0010991100003124

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022. https://doi.org/10.5555/944919.944937

[4] Philippe Caillou, Jonas Renault, Jean-Daniel Fekete, Anne-Catherine Letournel, and Michèle Sebag. 2021. Cartolabe: A Web-Based Scalable Visualization of Large Document Collections. *Computer Graphics and Applications* 41, 2 (2021), 76–88. https://doi.org/10.1109/MCG.2020.3033401

[5] Michael A. A. Cox and Trevor F. Cox. 2008. Multidimensional Scaling. In *Handbook of Data Visualization*. Springer, 315–347. https://doi.org/10.1007/978-3-540-33037-0_14

[6] Scott Deerwester, Susan T Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

[7] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. 2021. Toward a Quantitative Survey of Dimension Reduction Techniques. *Transactions on Visualization and Computer Graphics* 27, 3 (2021), 2153–2173. https://doi.org/10.1109/TVCG.2019.2944182

[8] Emden R. Gansner, Yifan Hu, and Stephen C. North. 2013. Interactive Visualization of Streaming Text Data with Dynamic Maps. *Journal of Graph Algorithms and Applications* 17, 4 (2013), 515–540. https://doi.org/10.7155/jgaa.00302

[9] T. Kohonen. 1997. Exploration of Very Large Databases by Self-organizing Maps. In *Proc. International Conference on Neural Networks (ICNN '97)*. IEEE, 1–6. https://doi.org/10.1109/ICNN.1997.611622

[10] Kostiantyn Kucher, Rafael M. Martins, and Andreas Kerren. 2018. Analysis of VINCI 2009–2017 Proceedings. In *Proc. 11th International Symposium on Visual Information Communication and Interaction (VINCI '18)*. ACM, 97–101. https://doi.org/10.1145/3231622.3231641

[11] Adrian Kuhn, David Erni, Peter Loretan, and Oscar Nierstrasz. 2010. Software Cartography: Thematic Software Visualization with Consistent Layout. *Journal of Software Maintenance and Evolution: Research and Practice* 22, 3 (2010), 191–210. https://doi.org/10.1002/smr.414

[12] Tuan M. V. Le and Hady W. Lauw. 2014. Semantic Visualization for Spherical Representation. In *Proc. 20th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 1007–1016. https://doi.org/10.1145/2623330.2623620

[13] Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv CoRR* stat.ML, 1802.03426 (2020), 63 pages. https://doi.org/10.48550/arXiv.1802.03426 pre-print.

[14] Rosane Minghim, Fernando Vieira Paulovich, and Alneu de Andrade Lopes. 2006. Content-based text mapping using multi-dimensional projections for exploration of document collections. In *Visualization and Data Analysis 2006*, Vol. 6060. SPIE, 259–270.

[15] F.V. Paulovich and R. Minghim. 2006. Text Map Explorer: a Tool to Create and Explore Document Maps. In *Tenth International Conference on Information Visualisation (IV'06)*. 245–251. https://doi.org/10.1109/IV.2006.104

[16] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A Method for Visualizing and Interpreting Topics. In *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces*. ACL, 63–70. https://doi.org/10.3115/v1/W14-3110

[17] A. Skupin. 2004. The World of Geography: Visualizing a Knowledge Domain with Cartographic Means. *Proc. National Academy of Sciences* 101, suppl 1 (2004), 5274–5278. https://doi.org/10.1073/pnas.0307654100

[18] Yee Teh and Sam Roweis. 2002. Automatic alignment of local representations. *Advances in neural information processing systems* 15 (2002).

[19] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008), 27 pages.

[20] James A Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. 1995. Visualizing the Non-visual: Spatial Analysis and Interaction with Information from Text Documents. In *Proc. Visualization 1995 Conference*. IEEE, 51–58. https://doi.org/10.1109/INFVIS.1995.528686